LA-UR-05-9439

*Approved for public release;
distribution is unlimited.*

Title:

THE BX PROJECT: FEDERATING AND MINING USAGE LOGS FROM LINKING SERVERS

Author(s):

Johan Bollen - LANL
Oren Beit-Arie - Ex Libris Inc, Boston, MA
Herbert Van de Sompel - LANL

Submitted to:

CNI Coalition for Networked Information
Dec. 5-6, 2005 Fall Task Force Meeting
Phoenix, Arizona

ABSTRACT

The bX Project: Federating and Mining Usage Logs from Linking Servers

Johan Bollen
Digital Library, Research & Prototyping Team
Los Alamos National Laboratory

Oren Beit-Arie
Chief Strategy Officer
Ex Libris Group

Herbert Van de Sompel
Technical Staff Member, Resarch Library
Los Alamos National Laboratory

The bX project aims at unleashing the power of usage information that is recorded on an ongoing basis by OpenURL-compliant linking servers. The project is a collaboration between the Digital Library Research & Prototyping Team of the Los Alamos National Laboratory and Ex Libris.

The bX project aims at unleashing the power of usage information that is recorded on an ongoing basis by OpenURL-compliant linking servers. The project is a collaboration between the Digital Library Research & Prototyping Team of the Los Alamos National Laboratory and Ex Libris.

Interesting things can be done with usage information that has been recorded for a specific information source from a digital library. The starting point of the bX project is that even more interesting things can be done with usage information that has been recorded by a linking server, because such a server can record activities across multiple OpenURL-enabled information sources of a specific digital library environment. As such, logs from a linking server are highly representative of the activities and preferences of a user population that requests services from that specific linking server. The bX project further recognizes that even more interesting things can be done with usage logs that are federated across multiple linking servers. Indeed, as a federation of linking servers grows in size, the federated usage log database becomes increasingly representative of the activities of the global scholarly user base.

The bX project currently focuses on two applications of linking server logs, at both the local and the federated level:

- Recommender services: A tool driven by usage information to help users discover related information.
- Metrics services: A toolset for mining usage information in order to obtain a variety of metrics. Certain metrics may be used to inform collection development decisions. The bX project, however, takes special interest in the derivation of metrics that could function as new indicators of scholarly quality.

The Project Briefing will discuss the standards-based architecture of the system that is under development, in addition to the concepts that underlie the Recommender and Metrics services. Furthermore, working prototypes of both services will be demonstrated.

The bX project:
Federating and Mining Usage Logs from Linking Servers

Johan Bollen [1], Oren Beit-Arie [2], and Herbert Van de Sompel [1]

[1] Digital Library Research & Prototyping Team
Research Library, Los Alamos National Laboratory
[2] Ex Libris Inc., Boston, MA

jbollen@lanl.gov , oren@exlibris-usa.com , herbertv@lanl.gov

Acknowledgement:
Marvin Pollard (CalState)
Nathan McFarland (LANL RL)

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

Outline

1. Problem statement
2. Analysis of local usage data
3. Towards federated usage data
4. Collaborating on the bX project
5. Mining federated usage data
6. What's Next
7. Conclusion

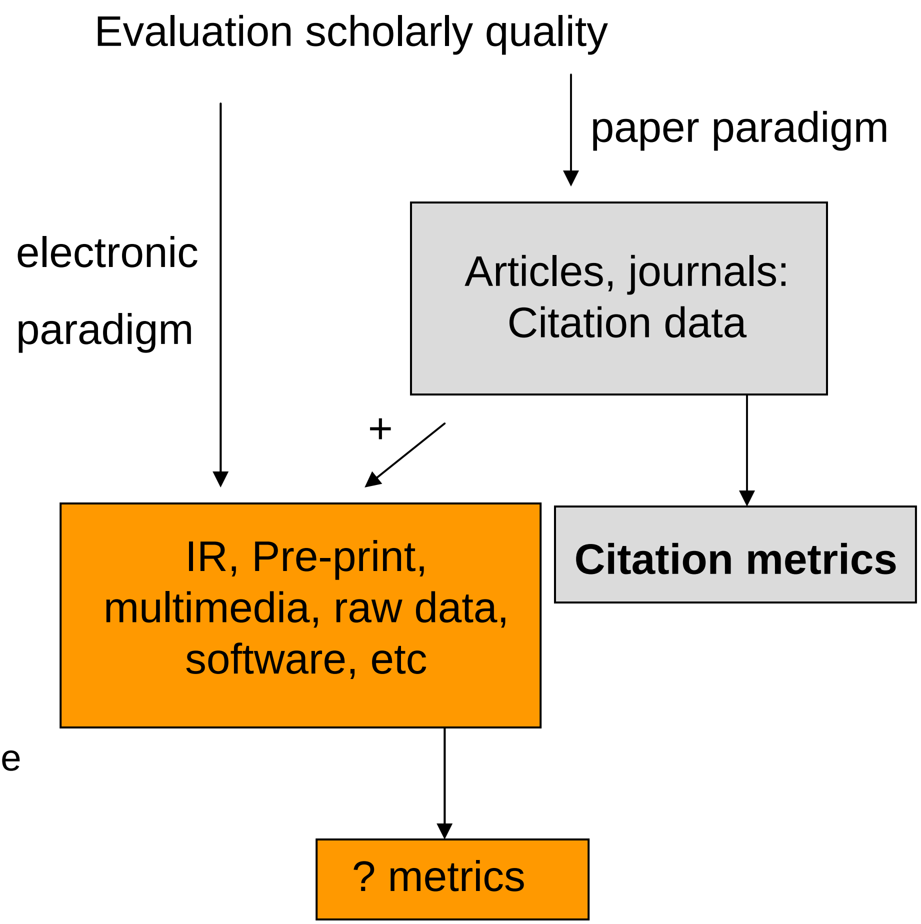
Outline

1. Problem statement
2. Analysis of local usage data
3. Towards federated usage data
4. Collaborating on the bX project
5. Mining federated usage data
6. What's Next
7. Conclusion

Scholarly evaluation in an electronic publishing paradigm

- Scholarly quality evaluated by citation counts
  - Domain: vetted literature only
  - Metrics: citation frequency
  - Limited resources: what and how we count

- Electronic paradigm changes everything
  - New models of communication:
    - Everything will be published
    - No central vetting authority
  - New models of scholarship
    - Publish multimedia, raw data, software
  - New metrics of evaluation?


Evaluation scholarly quality

paper paradigm

electronic paradigm

Articles, journals: Citation data

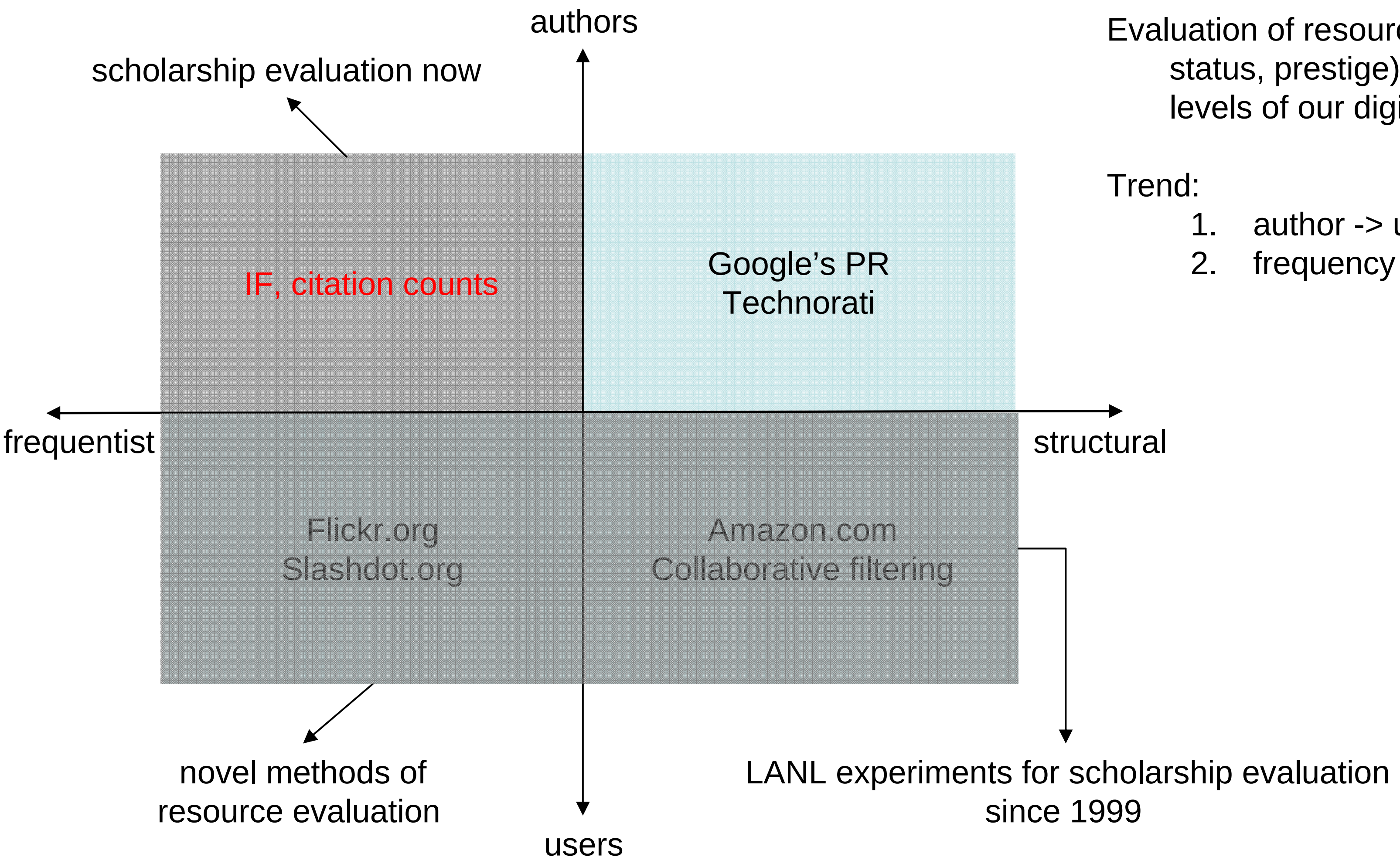
+

IR, Pre-print, multimedia, raw data, software, etc

Citation metrics

? metrics

Evaluation of resources: a user-driven revolution



authors

scholarship evaluation now

IF, citation counts

Google's PR
Technorati

frequentist

structural

Flickr.org
Slashdot.org

Amazon.com
Collaborative filtering

novel methods of
resource evaluation

users

LANL experiments for scholarship evaluation
since 1999

Evaluation of resources (quality, status, prestige) is required on all levels of our digital infrastructure.

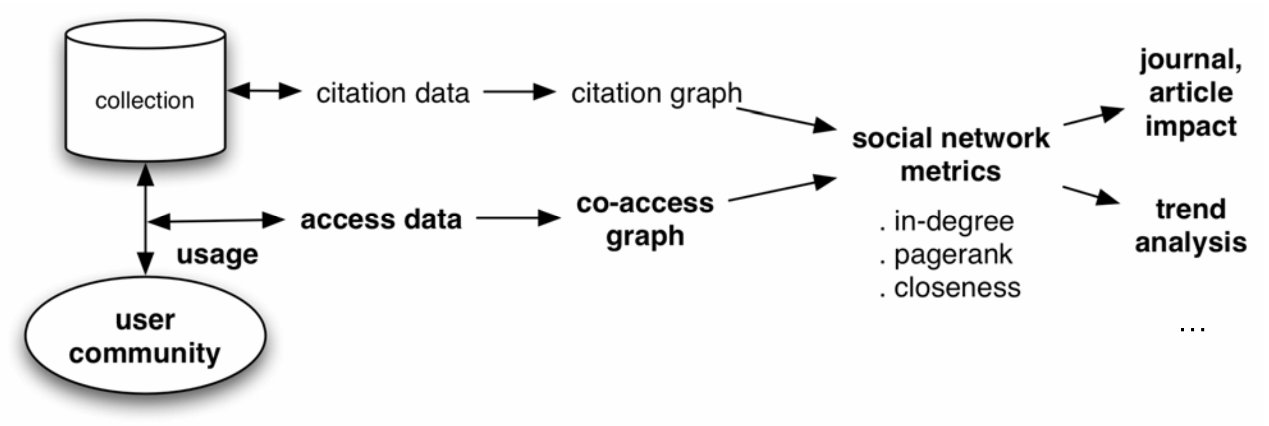
Trend:

1. author -> user
2. frequency -> structure

Outline

1. Problem statement
2. Analysis of local usage data
3. Towards federated usage data
4. Collaborating on the bX project
5. Mining federated usage data
6. What's Next
7. Conclusion

Scholarly evaluation: process flow for data analysis



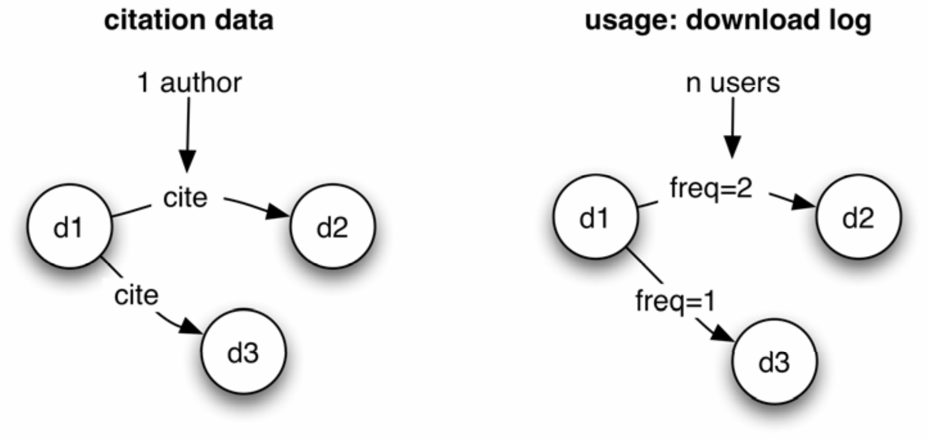
source ⟶ data ⟶ structure ⟶ metrics ⟶ evaluation

Usage: user activity that expresses interest or preference
Access data: particular instance(s) of usage (e.g. request abstract, download full-text)
Co-access: repeated instances of users accessing same pairs of items (documents)
Co-access graph: network of co-access data
Social network metrics: prestige from network structure

Scholarly evaluation: mining usage data and deriving metrics

Two essential components to move beyond descriptive usage stats:

1) **Datamine usage patterns for networks of items relationships**:
   - Citation: when A cites B, A and B are related
   - Usage: when A and B are frequently co-used, they are related



2) **Structural analysis of resulting networks**:
   - Social network metrics of visibility (in-degree), prestige (PageRank), power (betweenness), etc
   - Mapping techniques: multi-dimensional scaling, self-organizing maps

•Kothari (2003). On using page cooccurence …
•Kim (2004). A clickstream-based collaborative…
•Sarwar (2001. Item-based collaborative filtering

LANL experiments: demonstrating the power of usage data analysis

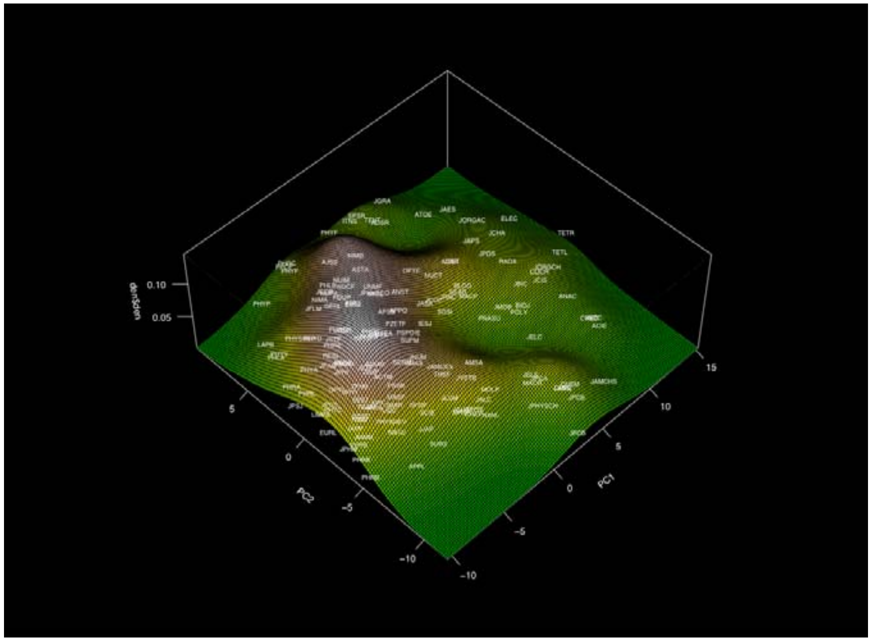
- LANL has been active in this area since early 1999
  - Early analysis of LANL RL usage data (local) in 1999
  - Extraction of item networks
  - Calculation of impact metrics (social network approach)
- Preliminary success
  - Demonstrated valid journal and article networks
  - Surprising success in ranking of items according to institutional focus
  - Discovery of hidden interest groups and focii
- Next two slides: recent results
  - February 2004 to April 2005
  - 392,455 usage events: any indication of preferences/interest
  - 5,866 users
  - 330,109 articles
  - 10,695 journals
- See publication list at end for more information

A comparison of 2004 LANL usage data and citation Impact Factor

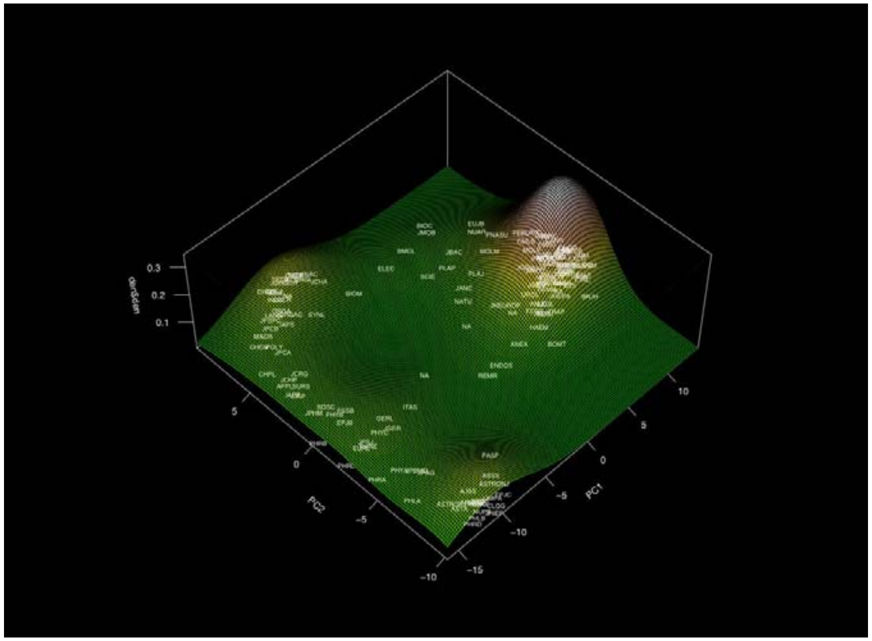
rank	Usage (PageRank)	IF (2003)	ISSN	Title
1	60.196	7.035	0031-9007	PHYS REV LETT
2	37.568	2.950	0021-9606	J CHEM PHYS
3	34.618	1.179	0022-3115	J NUCL MATER
4	31.132	2.202	1063-651X	PHYS REV E
5	30.441	2.171	0021-8979	J APPL PHYS
6	30.128	30.979	0028-0836	NATURE
7	29.972	29.781	0036-8075	SCIENCE
8	27.187	6.516	0002-7863	J AM CHEM SOC
9	24.602	4.049	0003-6951	APPL PHYS LETT
10	23.631	2.992	0148-0227	J GEOPHYS RES

Green: convergent
Red: divergent

Information landscapes



LANL 2004 Usage Data



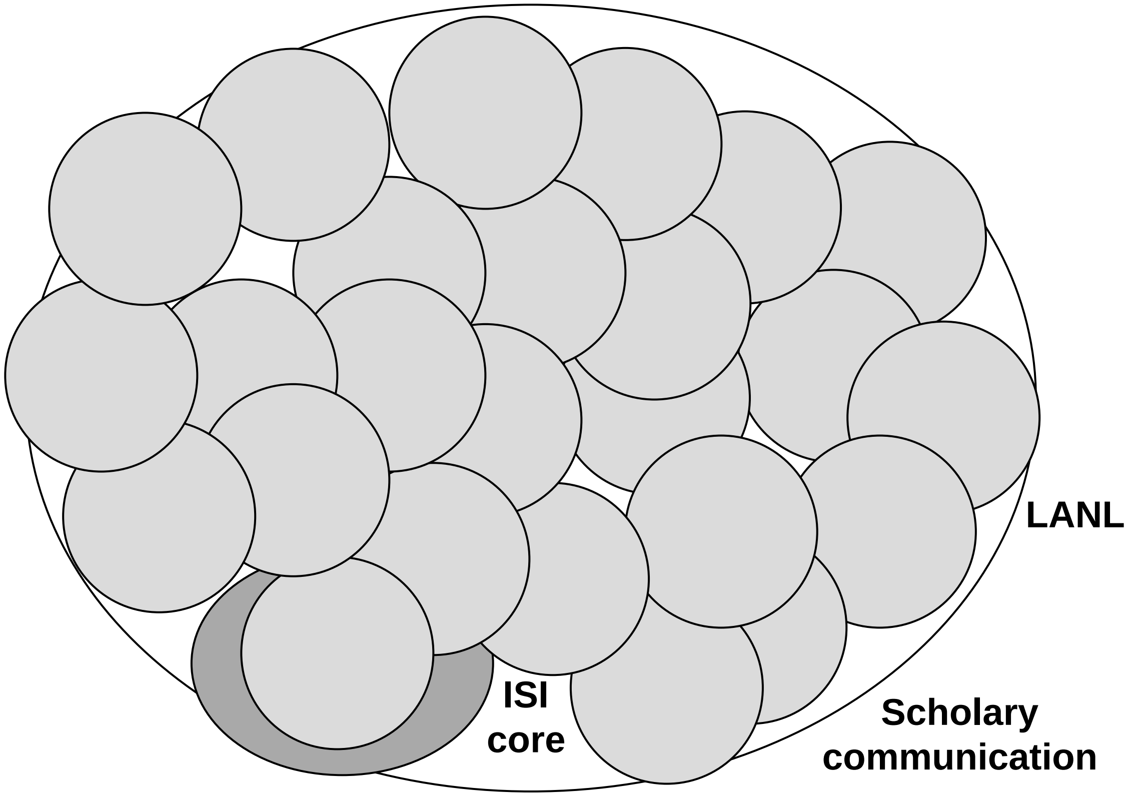
ISI Journal Citation Reports 2003

- Two component model
- Principal Component 1: Life vs. natural science
- Principal Component 2: Microscopic vs. macroscopic
- Z-axis: cluster density

Outline

1. Problem statement
2. Analysis of local usage data
3. Towards federated usage data
4. Collaborating on the bX project
5. Mining federated usage data
6. What's Next
7. Conclusion

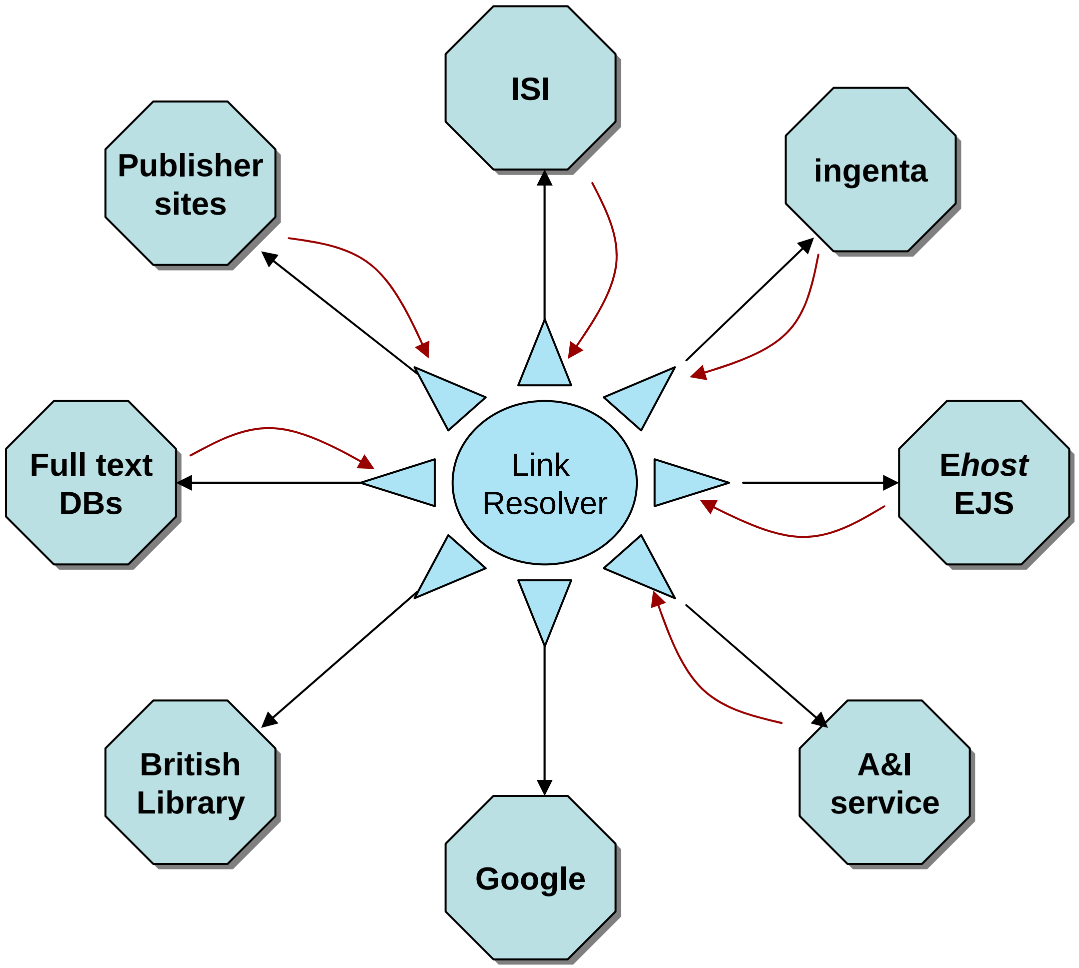
From local usage data to *global* usage data



- Local usage is interesting
  - Informs local collection management
  - Prominent communities can inform assessments of science trends
  - Covers wide range of communication items
  - Immediate availability

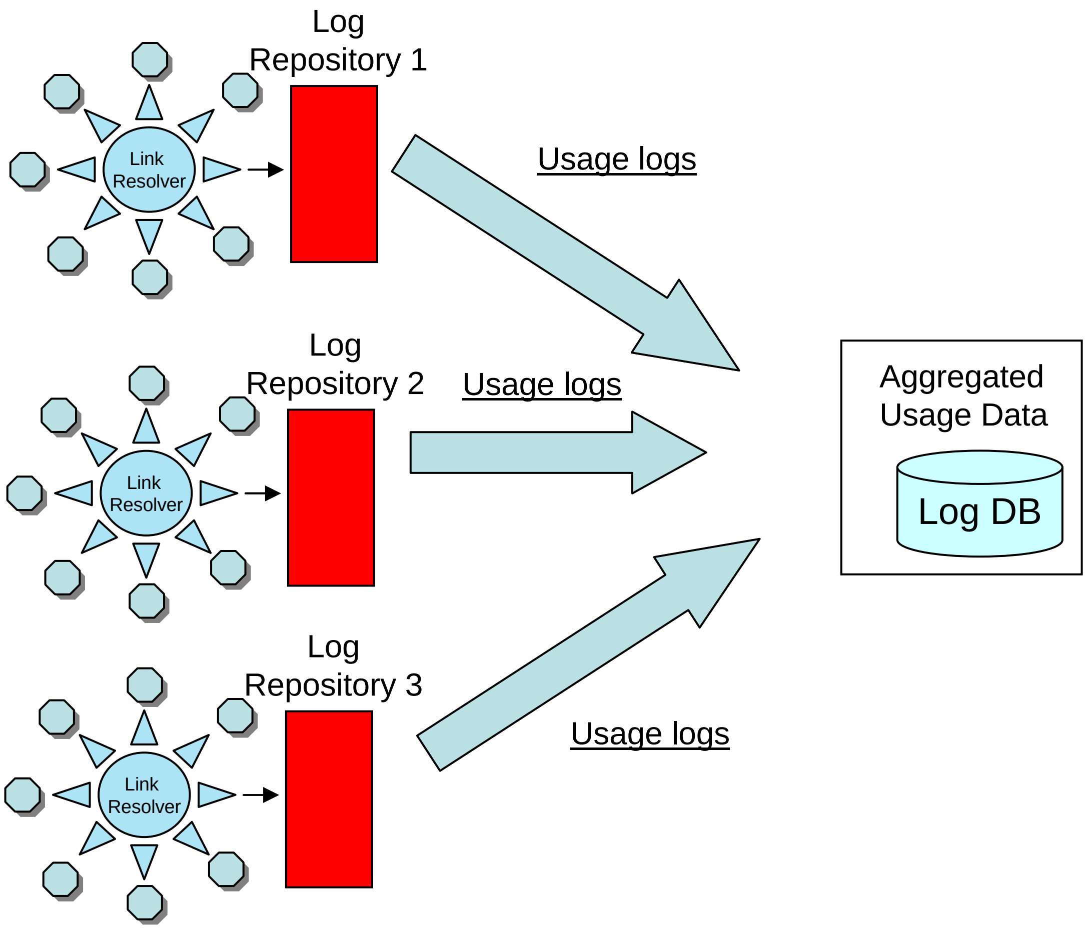
- Global, aggregated usage data is even more interesting
  - Monitor science as it takes place
  - Replace/augment/validate proprietary data sets
  - Allow free-form aggregation:
    - Clusters of institutions
    - Focus on sub-domains and communities

Local aggregation of usage data: linking servers



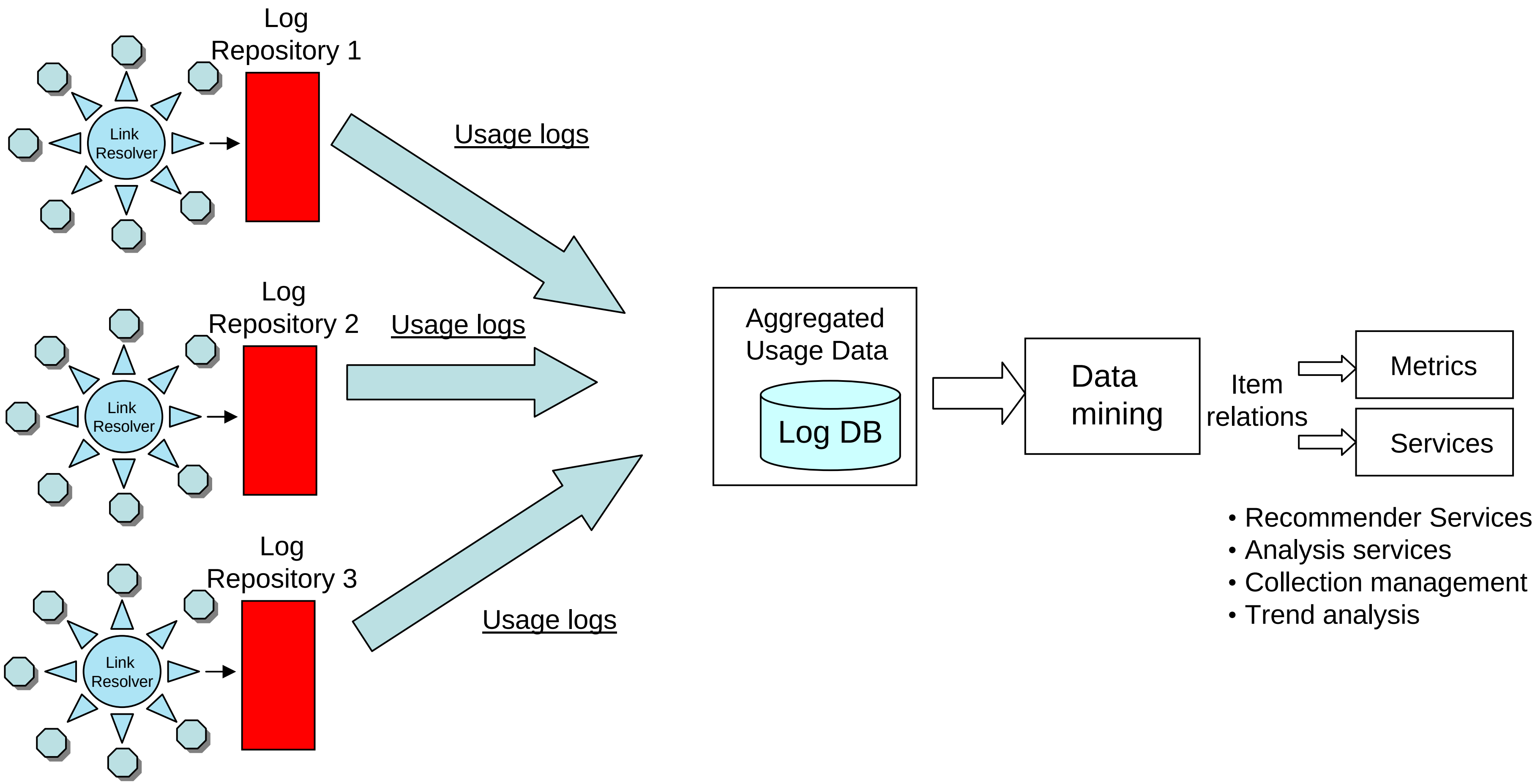
- Linking servers can record activities across multiple OpenURL-enabled information sources of a specific digital library environment
- Linking server logs are representative of the activities of a particular user population
- Global scholarly information space compliant with linking servers
- Allows recording of clickstream data: other methods of log aggregation can not connect “same user, different system” streams

Global aggregation of usage data



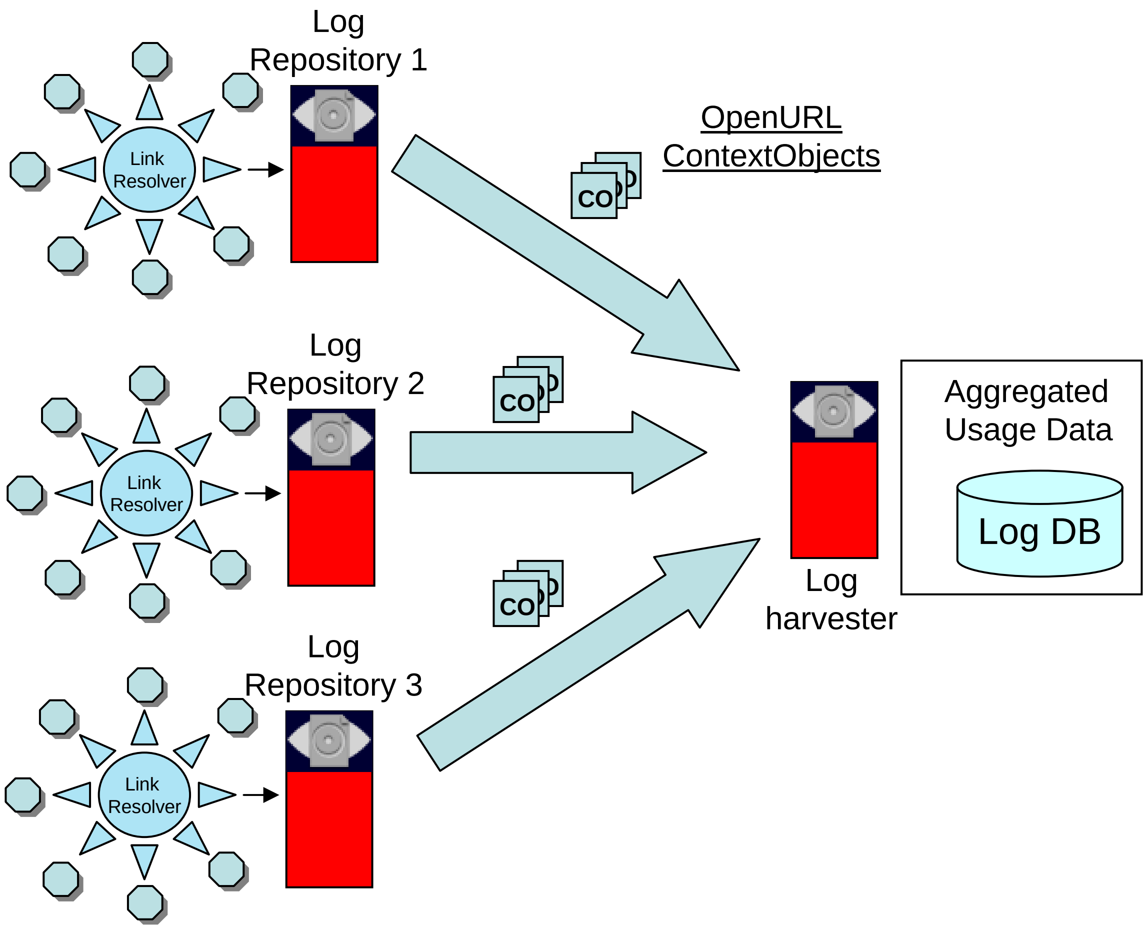
- Aggregation of linking server logs leads to data set representative of large sample of scholarly community
- Global really means different samples of scholarly community
    - Can be finetuned for local communities
    - Possibility of truly global coverage

Analysis and services based on *global* usage data



- Recommender Services
- Analysis services
- Collection management
- Trend analysis

bX project: standards-based aggregation of usage data

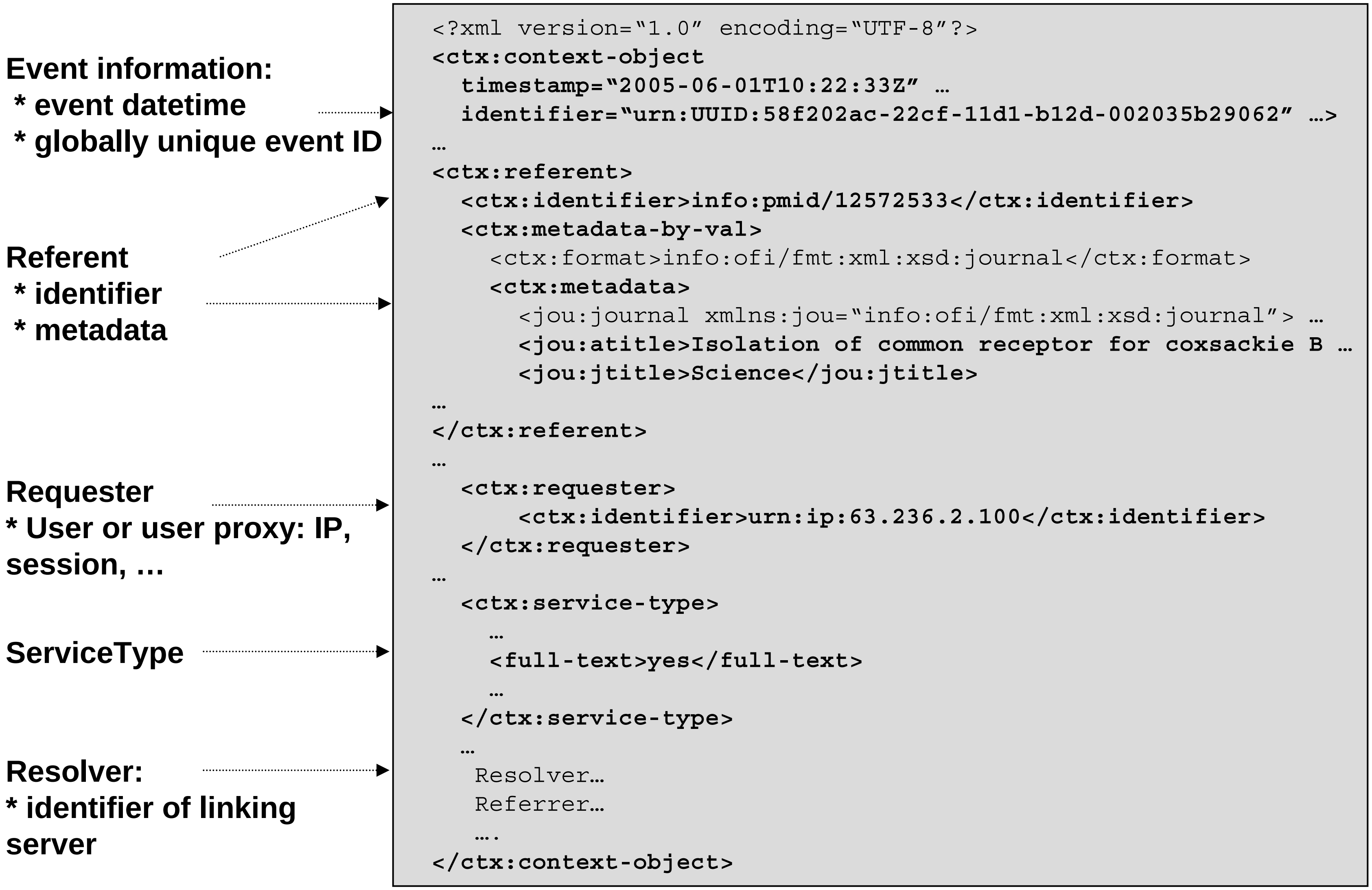


Usage log aggregation via OAI-PMH

Log Repository properties:

- OAI-PMH *metadata* record:
  - linking server event log for specific document in specific session
  - expressed using OpenURL XML ContextObject Format
- OAI-PMH identifier: UUID for event
- OAI-PMH datestamp: datetime the event was added to the Log Repository

bX project: OpenURL ContextObject to represent usage data



Event information:
*** event datetime**
*** globally unique event ID**

Referent
*** identifier**
*** metadata**

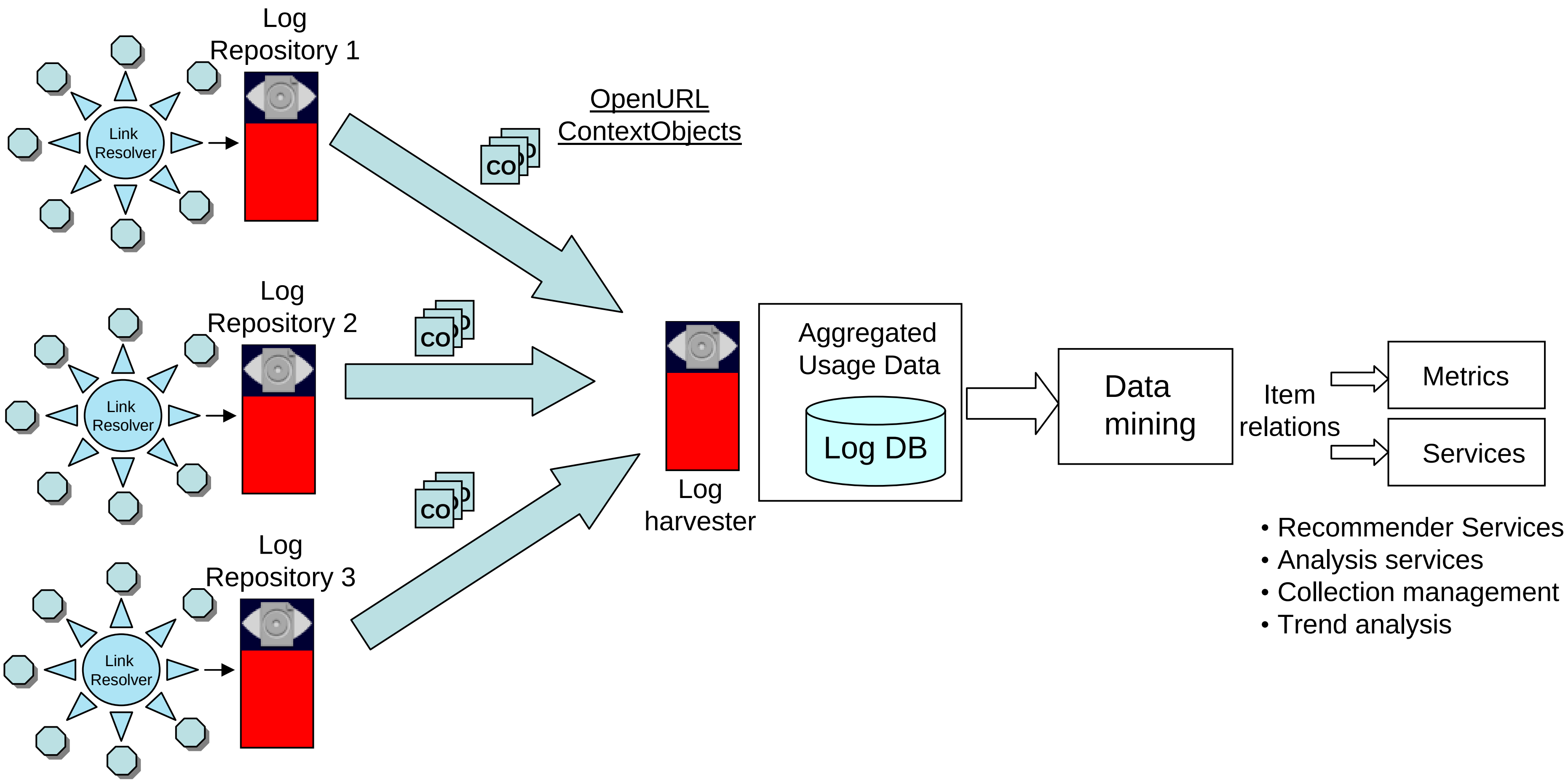
Requester
*** User or user proxy: IP, session, …**

ServiceType

Resolver:
*** identifier of linking server**

```
<?xml version="1.0" encoding="UTF-8"?>
<ctx:context-object
  timestamp="2005-06-01T10:22:33Z" …
  identifier="urn:UUID:58f202ac-22cf-11d1-b12d-002035b29062" …>
…
<ctx:referent>
  <ctx:identifier>info:pmid/12572533</ctx:identifier>
  <ctx:metadata-by-val>
    <ctx:format>info:ofi/fmt:xml:xsd:journal</ctx:format>
    <ctx:metadata>
      <jou:journal xmlns:jou="info:ofi/fmt:xml:xsd:journal"> …
      <jou:atitle>Isolation of common receptor for coxsackie B …
      <jou:jtitle>Science</jou:jtitle>
…
</ctx:referent>
…
  <ctx:requester>
      <ctx:identifier>urn:ip:63.236.2.100</ctx:identifier>
  </ctx:requester>
…
  <ctx:service-type>
    …
    <full-text>yes</full-text>
    …
  </ctx:service-type>
  …
   Resolver…
   Referrer…
   ….
</ctx:context-object>
```

bX project: analysis and services based on aggregated usage data



- Recommender Services
- Analysis services
- Collection management
- Trend analysis

bX project: analysis and services based on aggregated usage data

- Data mining:
  - Derive document relationships from access sequences
  - Use common techniques: clickstream datamining and association rule learning
- Metrics:
  - Recommender systems: item-based collaborative filtering and spreading activation
  - Common social network metrics of impact, prestige, prominence, etc

Outline

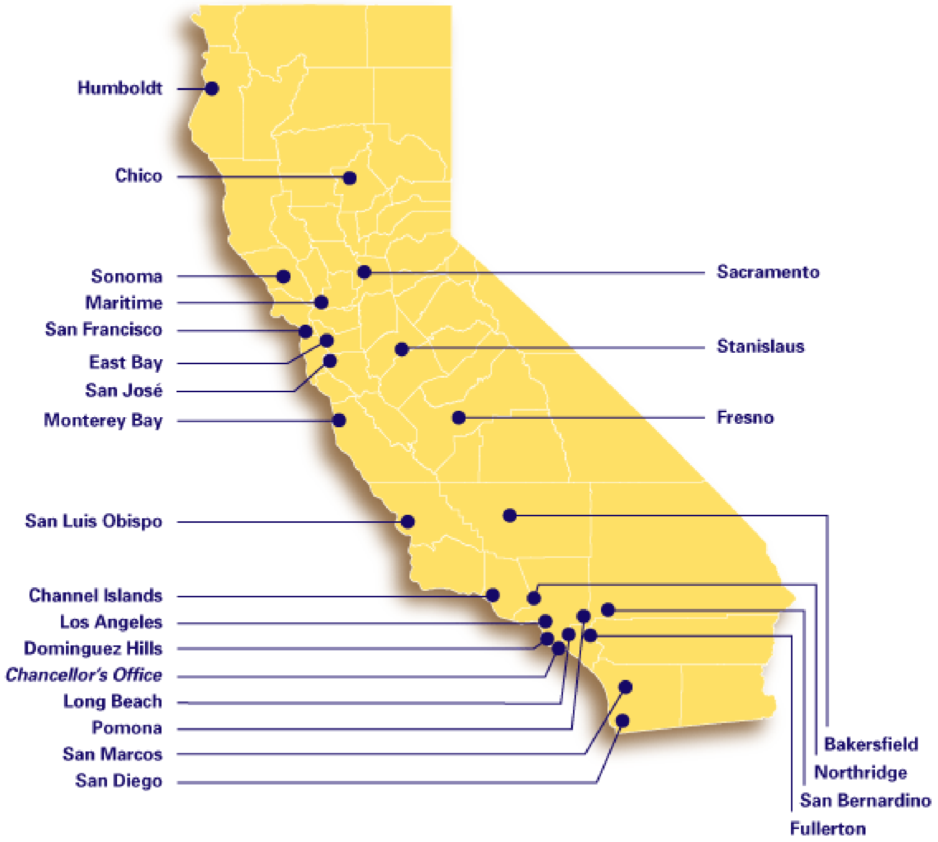
1. Problem statement
2. Analysis of local usage data
3. Towards federated usage data
4. **Collaborating on the bX project**
5. Mining federated usage data
6. What's Next
7. Conclusion

Partners and collaborations: Ex Libris/SFX

- Launched SFX in March 2001
- Co-developed the OpenURL
- About 900 libraries in 36 countries
  - 66% are members of consortia
  - 74 ARL libraries (60%)
  - Central and Local hosting
  - Growing usage
- Extensive usage logs
- Some relevant features:
  - Support for Z39.88-2004 (OpenURL 1.0)
    - SAP1 and SAP2
    - Internal representation of Context Object
  - Supports various consortia models
    - Supports distributive linking environments
- Involvement in bX:
  - Enabling role for research and development
  - Enhanced SFX to facilitate experimentation
  - Facilitate access to usage data sources

Partners and collaborations: CalState

- 23 campuses and seven off-campus centers,
- 409,000 students
- 44,000 faculty and staff
- SFX live since Fall 2002
- SFX consortium model: 23 instances (for each of the campuses) + 1 shared (the Chancellor’s Office, for shared resources)
- Involvement in bX: provided access to usage data for experimentation in framework of bX project



QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.

Outline

1. Problem statement
2. Analysis of local usage data
3. Towards federated usage data
4. Collaborating on the bX project
5. Mining federated usage data
6. What's Next
7. Conclusion

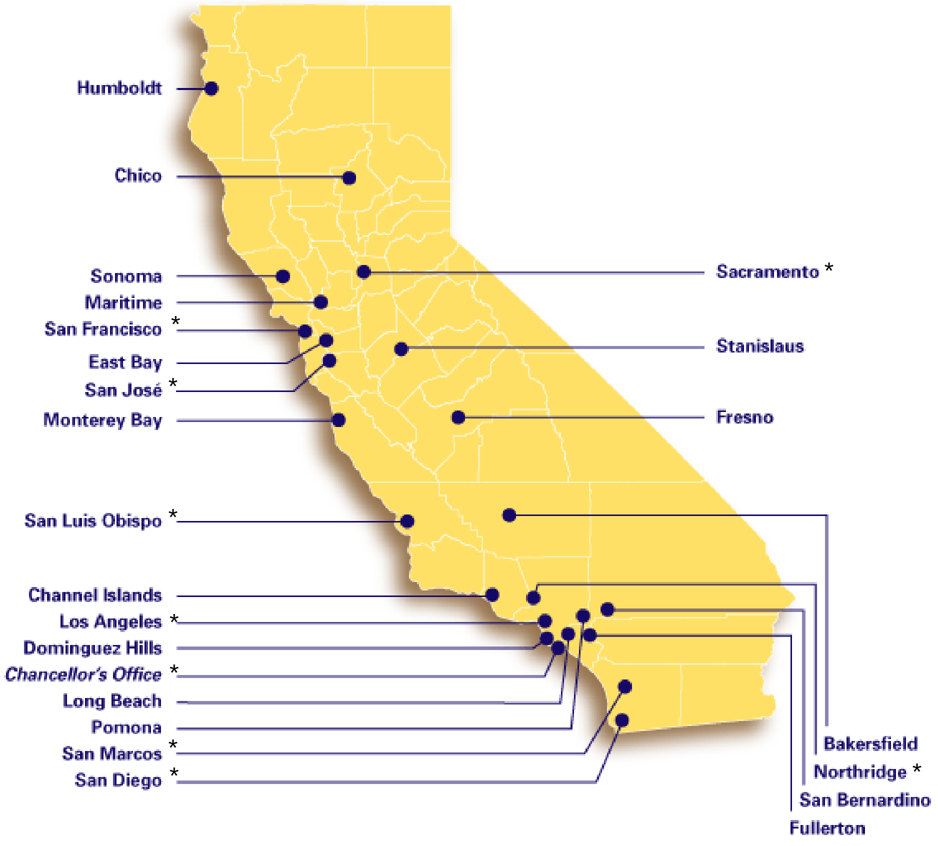
Mining federated usage data: CalState experiments

This is not pie in the sky: we have actually done it!

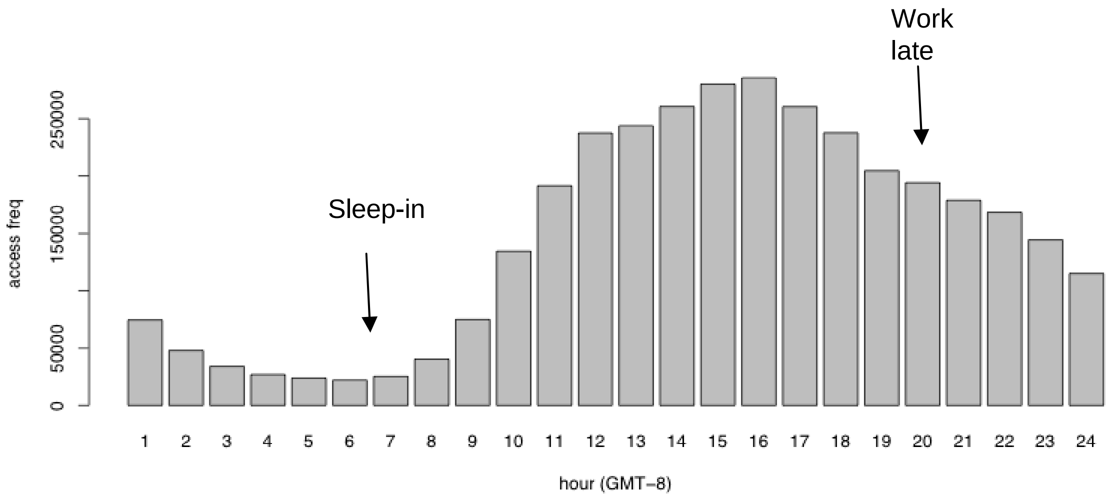
- Collaboration with CalState system via Ex Libris:
  - 23 campuses, seven off-campus centers, 409,000 students, and 44,000 faculty and staff
- CalState collaborator and point of contact:
  - Marvin Pollard (Chancellor’s office)

- Recorded usage includes all requests for which merged SFX menu has been presented:
  - Full-text requests
  - Abstract requests
  - Any expression of user interest

- **Present analysis covers 9 CalState institutions:**
  - **Chancellor, CPSLO, Los Angeles, Northridge., Sacramento, San Jose, San Marcos, SDSU, and SFSU**
  - **167,204 individuals, 3,507,484 accesses, 2,133,556 documents, Nov. 2003 - Aug. 2005**

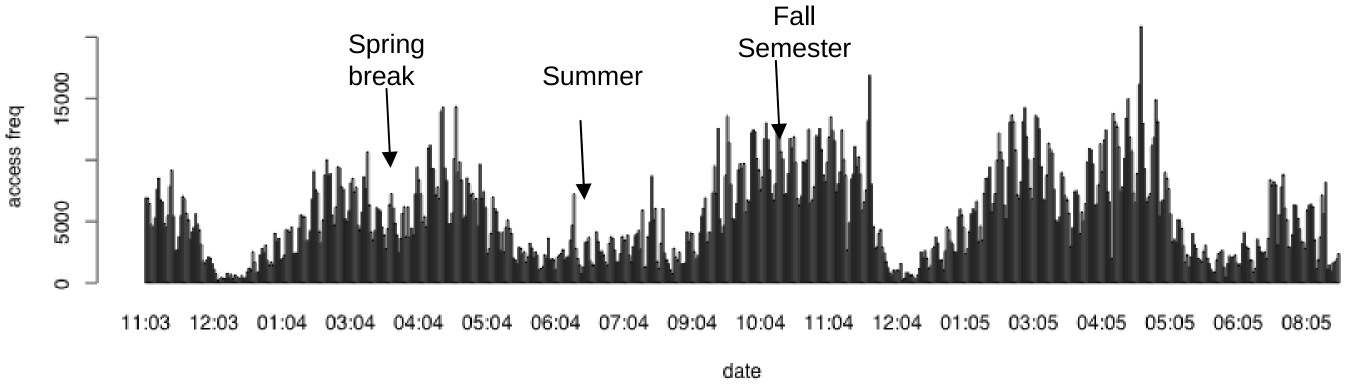


Some statistics: the academic rhythm


Work late

Sleep-in

QuickTime™ and a
Graphics decompressor
are needed to see this picture.


Spring break

Summer

Fall Semester

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

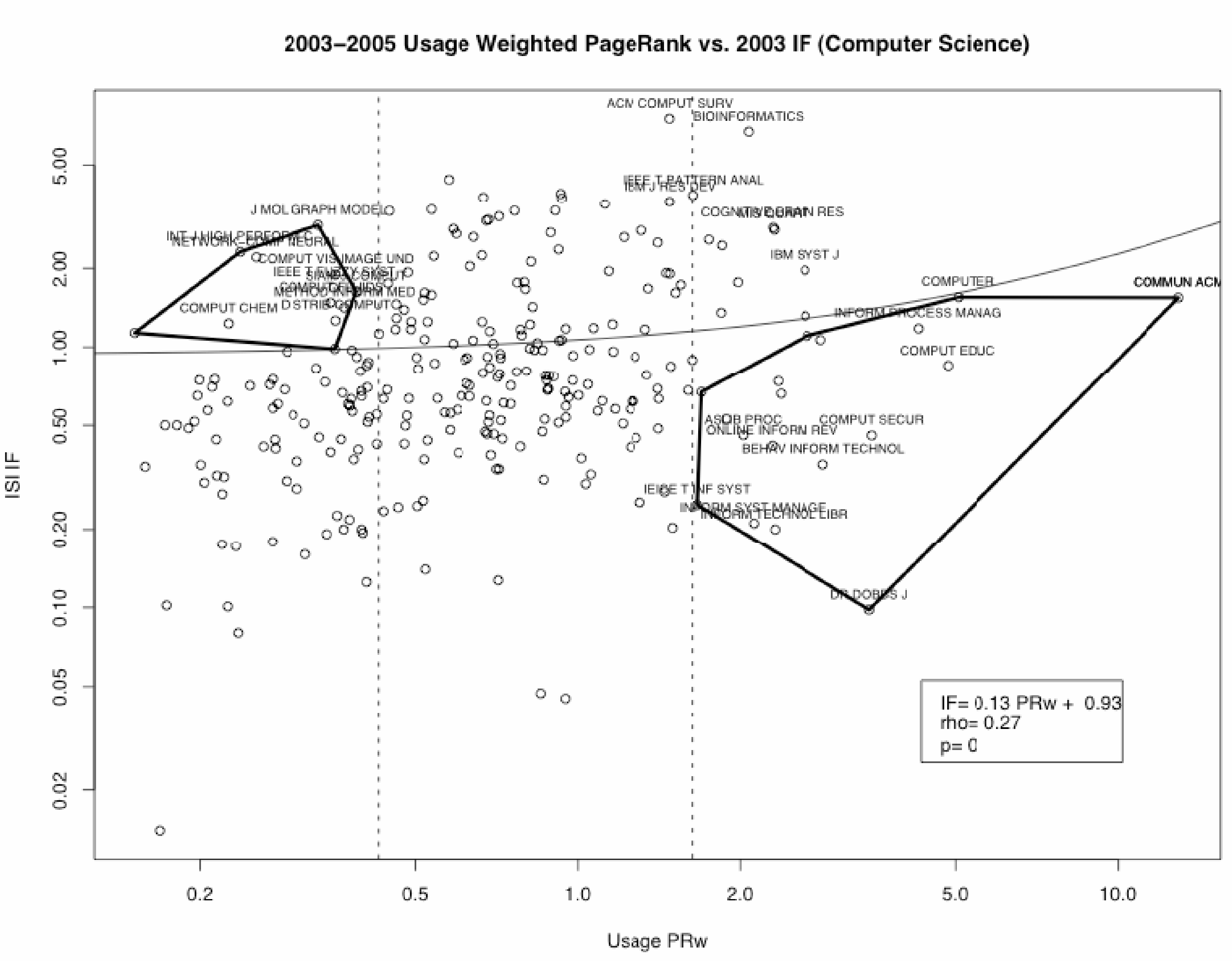
Results: journal ranking

rank	Usage (PageRank)	IF (2003)	ISSN	Title
1	78.565	21.455	0098-7484	JAMA-J AM MED ASSOC
2	71.414	29.781	0036-8075	SCIENCE
3	60.373	30.979	0028-0836	NATURE
4	40.828	3.779	0890-8567	J AM ACAD CHILD PSY
5	39.708	7.157	0002-953X	AM J PSYCHIAT
6	38.113	34.833	0028-4793	NEW ENGL J MED
7	37.492	3.363	0090-0036	AM J PUBLIC HEALTH
8	37.031	2.591	0195-9131	MED SCI SPORT EXER
9	27.248	0.998	0309-2402	J ADV NURS
10	26.987	5.692	0002-9165	AM J CLIN NUTR

Green: convergent
Red: divergent

QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.

Comparison of journal usage PageRank and citation Impact Factor

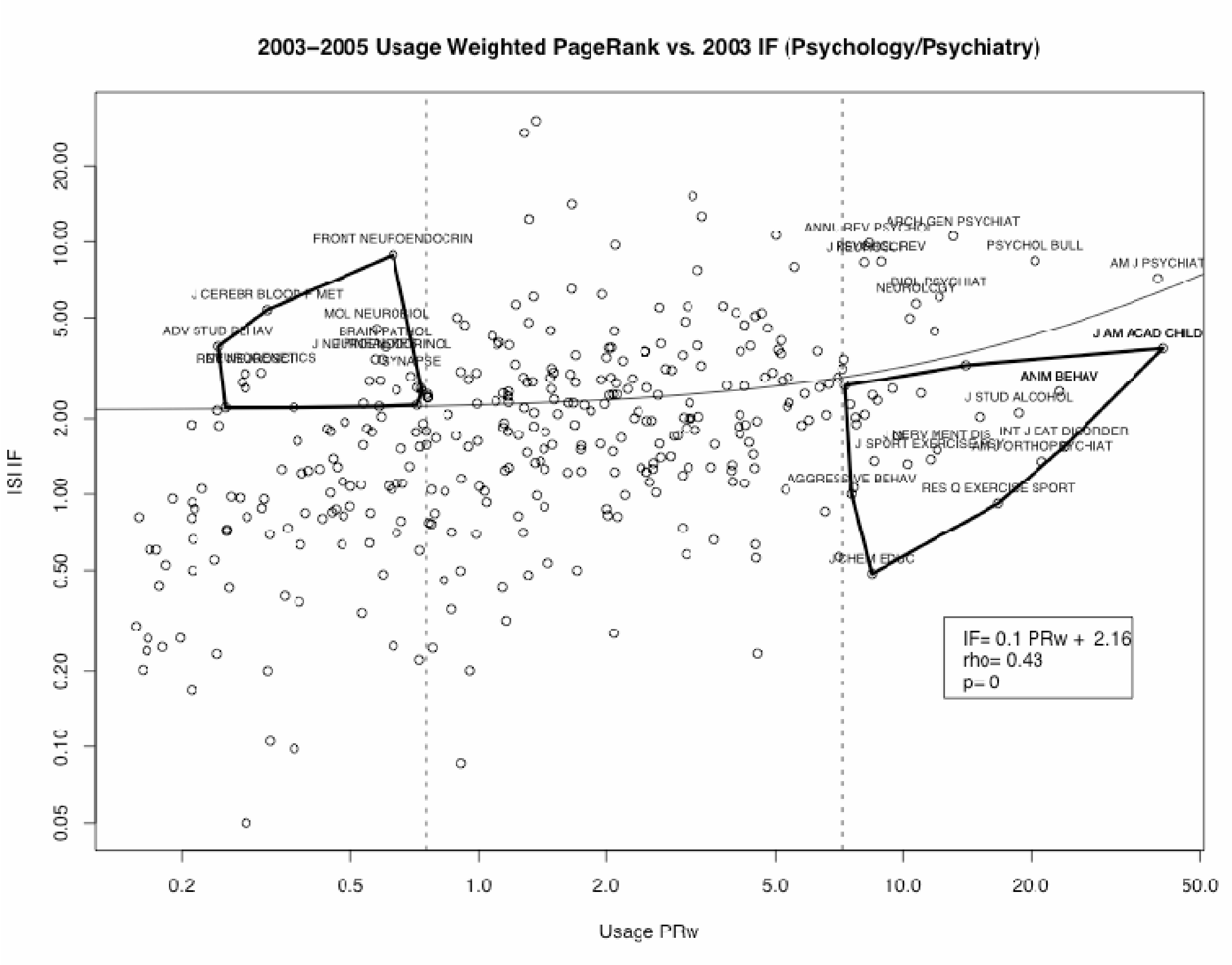


COMPUTER SCIENCE

QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.

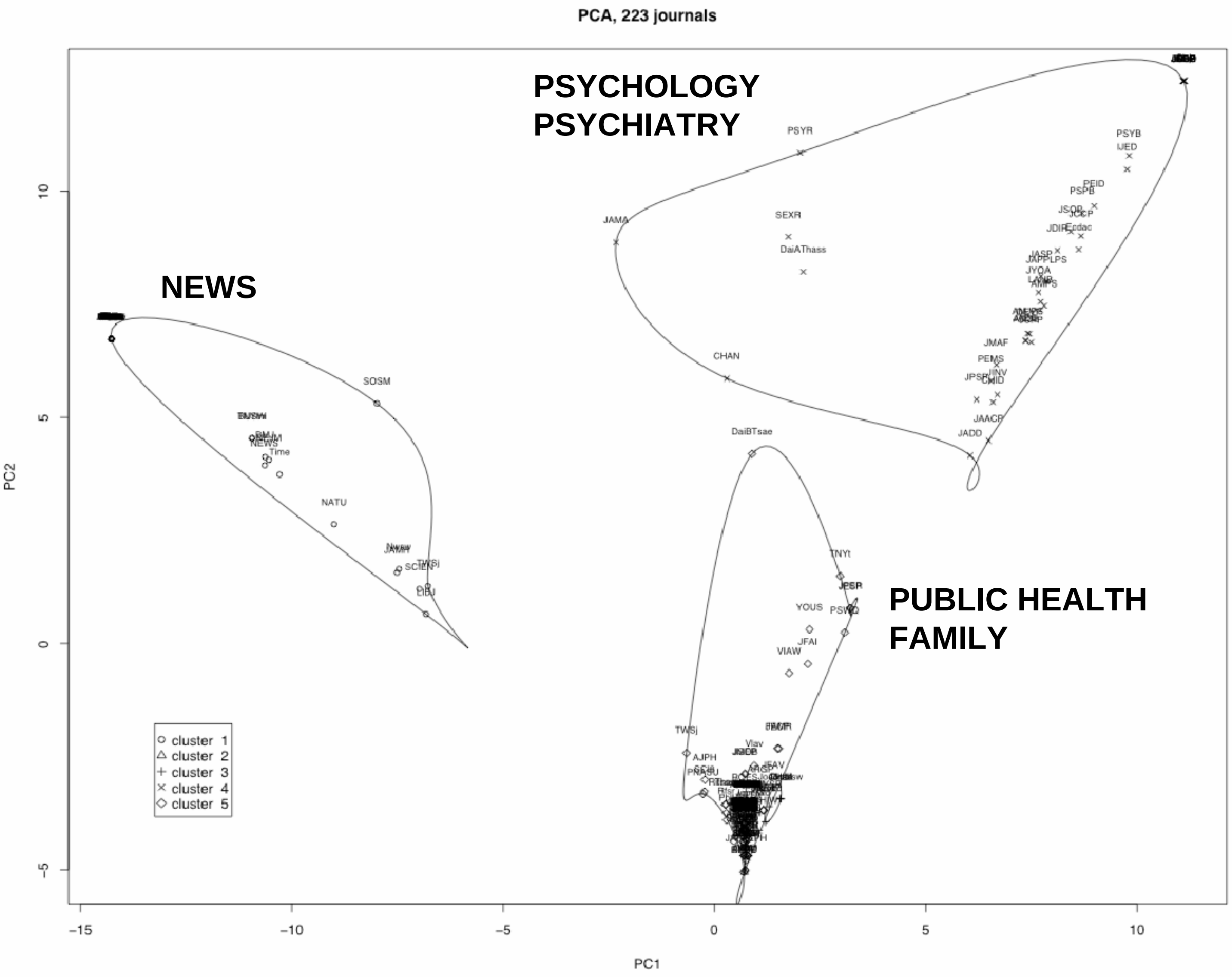
The bX Project: Federating and Mining Usage Logs from Linking Servers
Johan Bollen, Oren Beit-Arie, Herbert Van de Sompel
CNI Fall 2005, December 5th - 6th 2005, Phoenix, Arizona, USA

Comparison of journal usage PageRank and citation Impact Factor



PSYCHOLOGY
PSYCHIATRY

Mapping the structure of science



PSYCHOLOGY
PSYCHIATRY

NEWS

PUBLIC HEALTH
FAMILY

QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.

Usage-based recommender system

- Operates on network derived from aggregated usage data
- Starts from (set of) documents (articles or journals)
- Scans usage network links for directly and indirectly related documents
- Results:
  - Scalable
  - Highly efficient
  - Highly relevant results derived from accumulated, aggregated usage data

Movie: article level recommendations

Movie: journal level recommendations

Outline

1. Problem statement
2. Analysis of local usage data
3. Towards federated usage data
4. Collaborating on the bX project
5. Mining federated usage data
6. What's Next
7. Conclusion

General issues

- **Privacy and other legal issues involved in large-scale usage recording**: user and session identification, legal implications of log storage, ownership, retention policies
- **Data validity**: usage definition, recording and representation, quality benchmarks, falsification issues
- **Metrics:** frequency, structure, mappings and trends
- **Aggregation and scalability:**
  - different architectural frameworks: linking server-based, other, scalability, anonymization issues
  - social/economic models of aggregation: trusted log repository, incentives, sampling issues
- **Log data processing:**
  - Datamining approaches: support from informetric and bibliometric community, Grouping, isolating and aggregating useful usage patterns
  - Cross-validation issues: comparison and validation to citation data, data validity metrics
- **Metrics and services:** informetric indicators, interfaces with existing bibliometric products, definition of end-user services
- **Advocacy, strategies and policies: implications for IR and OA movement**

What’s next?

- Emerging activities in the realm of applications of usage data :
  - Mellon Foundation workshop on Usage Data, early 2005
  - DINI meeting Humboldt-Universität zu Berlin
  - SUSHI: Standardized Usage Statistics Harvesting Initiative (Harvard, Thomson Scientific, Cornell, and others)
  - IRS: Interoperable Repository Statistics (U. Southampton)

- LANL and Ex Libris exploring further collaboration in the realm of bX

Outline

1. Problem statement
2. Analysis of local usage data
3. Towards federated usage data
4. Collaborating on the bX project
5. Mining federated usage data
6. What's Next
7. Conclusion

Conclusion

- Scholarly communication is going through a revolution
- Scholarly evaluation will too! Focus will be on
  - Immediacy
  - Representativeness
  - Openness, standards and scalability
  - Acknowledging structural aspects of prestige and impact in the scholarly community
- User driven evaluation offers an interesting alternative to current short-front evaluation methods in a long-tail world

- Feasibility of usage analysis demonstrated at local and *global* level
  - LANL results indicate:
    - Possibility of local prestige and impact ranking
    - Additional usage-based services such as recommender systems possible
  - bX project on aggregated data and analysis:
    - Large-scale aggregation demonstrated scalability
    - Use of existing standards ensures openness, ability of all to participate
    - Possibility of spontaneous emergence of vetting and standardization system for usage quality indicators

Some papers:

- **J. Bollen, H. Van de Sompel, J. Smith, and R. Luce**. Toward alternative metrics of journal impact: a comparison of download and citation data. *Information Processing and Management*, 41(6):1419-1440, 2005.
  - http://dx.doi.org/10.1016/j.ipm.2005.03.024
- **J. Bollen, R. Luce, S. Vemulapalli, and W. Xu**. Detecting research trends in digital library readership. In *Proceedings of the Seventh European Conference on Digital Libraries (LNCS 2769)*, pages 24-28, Trondheim, Norway, August 18 2003. Springer-Verlag.
  - http://www.springerlink.com/openurl.asp?genre=article&issn=0302-9743&volume=2769&spage=24
- **J. Bollen, R. Luce, S. Vemulapalli, and W. Xu**. Usage analysis for the identification of research trends in digital libraries. *D-Lib Magazine*, 9(5), 2003.
  - http://www.dlib.org/dlib/may03/bollen/05bollen.html

The bX project: Federating and mining usage logs from linking servers

Johan Bollen (LANL), Oren Beit-Arie (Ex Libris) and Herbert Van de Sompel (LANL).

SUMMARY: The bX project aims at unleashing the power of usage information that is recorded on an ongoing basis by OpenURL-compliant linking servers. The project is a collaboration between the Digital Library Research & Prototyping Team of the Los Alamos National Laboratory and Ex Libris.

Interesting things can be done with usage information that has been recorded for a specific information source from a digital library. The starting point of the bX project is that even more interesting things can be done with usage information that has been recorded by a linking server, because such a server can record activities across multiple OpenURL-enabled information sources of a specific digital library environment. As such, logs from a linking server are highly representative of the activities and preferences of a user population that requests services from that specific linking server. The bX project further recognizes that even more interesting things can be done with usage logs that are federated across multiple linking servers. Indeed, as a federation of linking servers grows in size, the federated usage log database becomes increasingly representative of the activities of the global scholarly user base.

RATIONALE: The evaluation of science is still largely based on citation and authorship data which originates in a paper-based process of communication, i.e. journals publishing a limited selection of approved articles. This model, exemplified by the Institute for Scientific Information's Impact Factors and Journal Citation Reports, is rapidly being made obsolete by advances in the evaluation of web resources which favor large-scale usage over expert opinion and structural metrics over voting procedures. Usage data combined with structural Google-type metrics of quality forms a powerful combo for the evaluation of scholarly communication items in a future where the predominant model of scholarly communication will be user-driven, decentralized and focused on a wide variety of item types, e.g. data sets, software and educational resources in addition to journal articles.

Fig. 1. shows a taxonomy which classifies evaluation techniques on the basis of whether the data they use is authored, e.g. citation and hyperlinks, versus user-generated, e.g. ratings and purchase patterns, and whether the metrics used are frequentist, e.g. citation counts, versus structural, e.g. Google's PageRank. The taxonomy shows how the present evaluation of scholarly communication items remains confined to the use of frequentist metrics applied to citation data, i.e. a publication's popularity among a group of author experts as indicated as its citation frequency, whereas most advances relevant to emerging publishing and research models have been made in the other 3 quadrants.

APPROACH: The bX project has developed and implemented an architecture which allows the large-scale aggregation of logs generated by openURL-enabled linking servers and their subsequent analysis for the evaluation of scholarly communication items. Fig. 2 and 3 show more details.

The architecture functions according to the following three phases:

1) Linking server logs are serialized as XML-ized OpenURL ContextObjects and exposed by an OAI-PMH repository. The repository retains full control of *what* is exposed and *how*, e.g. anonymization of user IDs. A trusted third-party (or federation thereof) can harvest and aggregate logs from a range of repositories thereby creating a usage data set representative of a particular community which in principle could extend to the entire scholarly community.

2) Next the aggregated logs are subjected to datamining techniques which derive item networks from access sequences recorded in the logs under the assumption that similar items are accessed by similar users. These networks can be used to construct recommender services useful to both local institutions as well as third-party aggregators.

3) As a last step, structural metrics of quality can be derived from the generated item networks leading to more complete, fine-grained and reliable evaluation of scholarly communication. Log data furthermore is free from publication delays and can be used to track immediately contemporary trends in science.

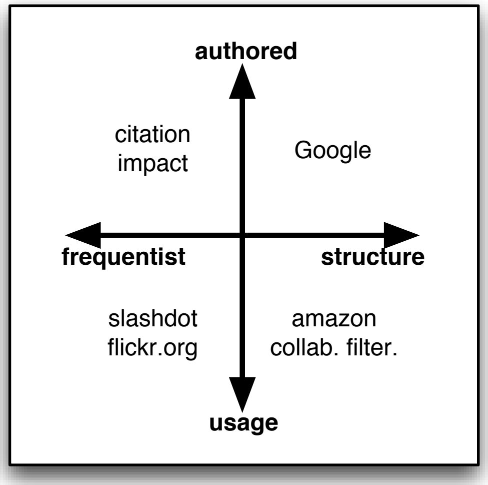
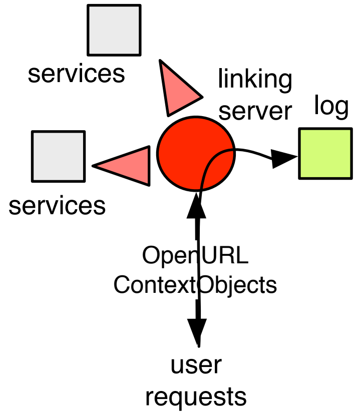


Fig. 1: Taxonomy of resource evaluation

CNI 2005
Phoenix, AZ

The bX project: Federating and mining usage logs from linking servers

Johan Bollen (LANL), Oren Beit-Arie (Ex Libris) and Herbert Van de Sompel (LANL)



services

linking server

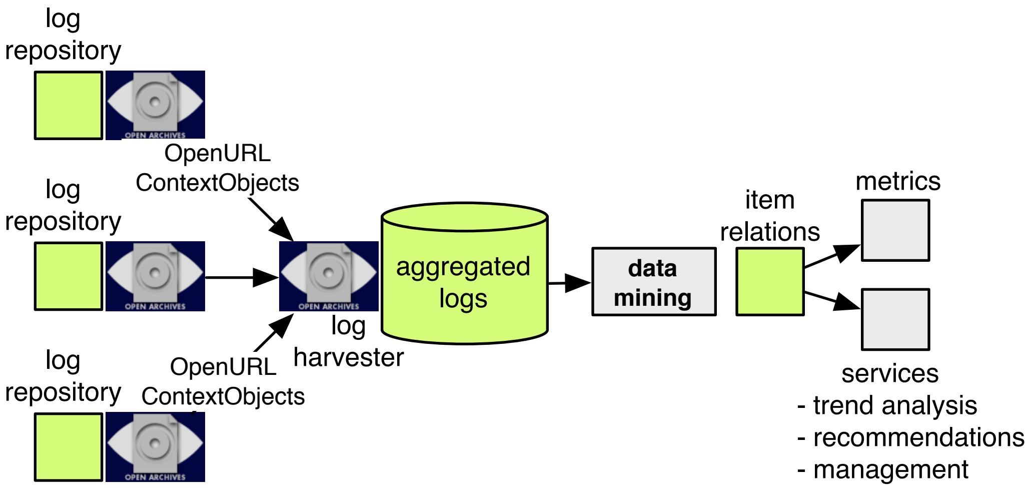
log

services

OpenURL ContextObjects

user requests

Fig. 2: Linking servers store OpenURL ContextObject logs.



log repository

log repository

log repository

OpenURL ContextObjects

OpenURL ContextObjects

log harvester

aggregated logs

data mining

item relations

metrics

services

- trend analysis
- recommendations
- management

Fig. 3: System architecture for the federating and mining of usage logs.

RESULTS: In collaboration with Ex Libris and CalState, the bX project has recently worked on aggregated usage data collected at over 20 academic institutions using the SFX link resolver. The particular data set, recorded in Nov. 2004 to Aug. 2005, concerns 167,204 individuals and 3,507,484 accesses. The collected usage data was analyzed with two principal objectives in mind. First, we analyzed the aggregated usage data resulting in the generation of quality indicators of scholarly communication items. Second, we prototyped a recommender service aimed at assisting the user in discovering related resources. Fig. 4. shows a mapping of science domains as produced from 2004-2005 LANL usage data. Fig. 5. shows a comparison of usage PageRank at CalState to the 2003 ISI IFs indicating significant overlap but interesting deviations. Our presentation will discuss promising results in both the area of science metrics and recommender systems.

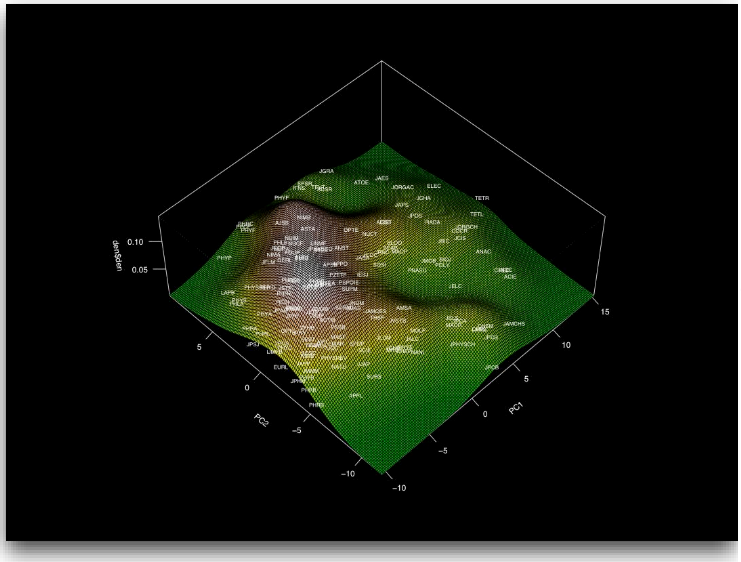


Fig. 4: Usage landscape reveals interdisciplinary clusters of research interests

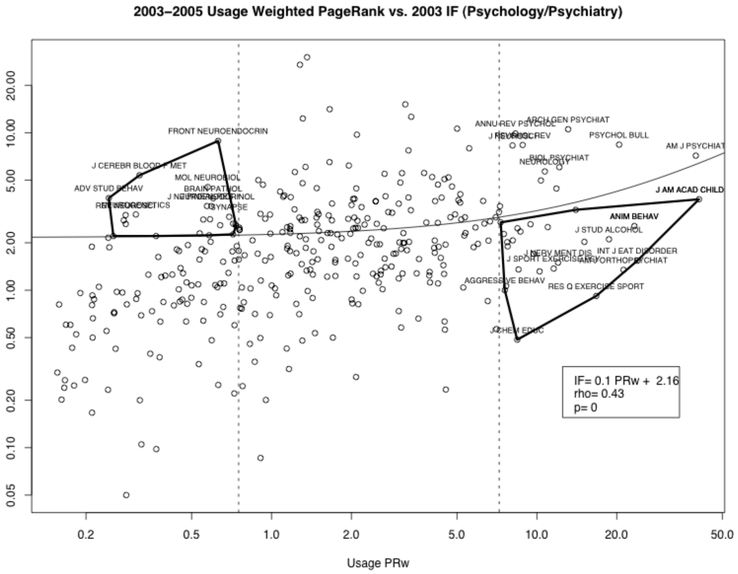


Fig. 5: Scatterplot of usage PageRank versus 2003 ISI IF

CNI 2005
Phoenix, AZ